Introduction to Statistical Analysis

BigData Week4

2025. 3.27

Eunhui Kim (김은희)
ehkim@kisti.re.kr

Contents

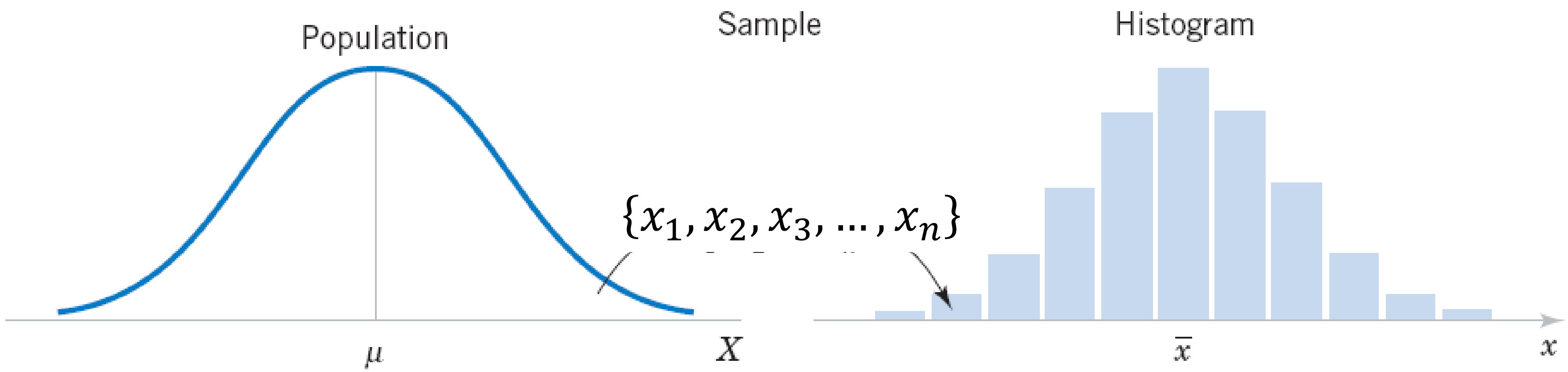
Understanding data by statistical Analysis

- The difference between ideal and real
  - The data we want to know : population(모집단),
  - The data we have : sample (표본)

- The field of statistical inference consists of those methods used to make decisions or draw conclusions about a **population**.

- These methods utilize the information contained in a **sample** from the population in drawing conclusions.

Understanding data by statistical Analysis



Relationship between a population and a sample.

μ, *population average*
σ, *population standard deviation*

$\bar{x}$, *sample average*
s, *sample standard deviation*

Understanding data by statistical Analysis

μ, *population average*
σ, *population standard deviation*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

x_i represents each value in the population.
N is the number of values in the population.

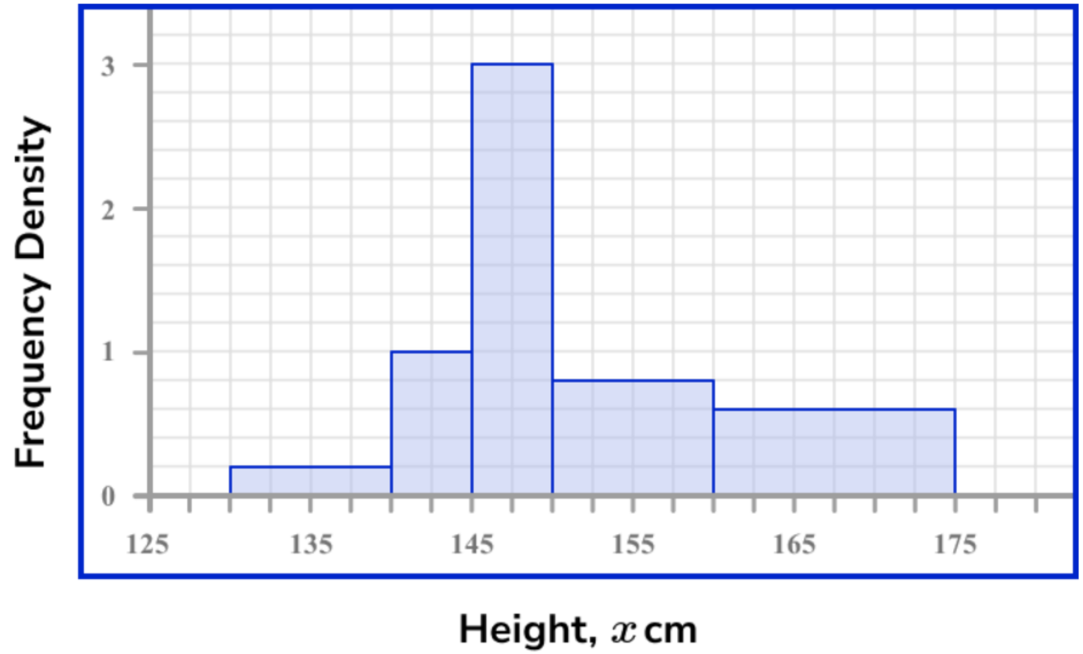
$\bar{x}$, *sample average*
s, *sample standard deviation*

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$\bar{x}$ is the sample mean.
x_i represents each value in the sample.
n is the number of values in the sample.

Understanding data by statistical Analysis

Height, cm	Frequency	Frequency Density
$130 \leq x < 140$	2	0.2
$140 \leq x < 145$	5	1
$145 \leq x < 150$	15	3
$150 \leq x < 160$	8	0.8
$160 \leq x < 175$	9	0.6



Estimation & Significance Testing

❖ Estimation (of population parameters)

Ex. "Based on GSS data, we're 95% confident that the population mean of the variable LONELY (no. of days in past week you felt lonely, $\bar{y}$ = 1.5, s = 2.2) falls between 1.4 and 1.6.

❖ Significance Testing

(Making decisions about hypotheses regarding "effects" and associations)
Ex. Article in Science 2008:
"We hypothesized that spending money on other people
has a more positive impact on happiness than spending money on oneself"

Statistical Inference : Estimation

Goal: How can we use sample data to estimate values of population parameters?

❖ **Point estimate**:
A single statistic value that is the "best guess" for the parameter value

❖ **Interval estimate**
An interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value.
Called a ***confidence interval***.
(Based on sampling distribution of the point estimate)

What is Random Variables?

- A **variable** is any characteristic, observed or measured. A variable can be either **random** or **constant** in the population of interest.

- For a defined population, every **random variable** has an associated distribution that defines the **probability** of occurrence of each possible value of that variable (if there are a finitely countable number of unique values) or all possible sets of possible values (if the variable is defined on the real line).

Probability Distribution?

A **probability distribution** (function) is a list of the probabilities of the values (simple outcomes) of a random variable.

Table: Number of heads in two tosses of a coin

y *outcome*	P(y) *probability*
0	1/4
1	2/4
2	1/4

For some experiments, the probability of a simple outcome can be easily calculated using a specific **probability function.** If y is a simple outcome and p(y) is its probability.

$$0 \leq p(y) \leq 1$$

$$\sum_{\text{all } y} p(y) = 1$$

Discrete Distributions

Relative frequency distributions for “counting” experiments.

- Bernoulli Distribution
- Binomial Distribution
- Negative Binomial
- Poisson Distribution
- Geometric Distribution
- Multinomial Distribution

Yes-No responses.

Sums of Bernoulli responses

Number of trials to kth event

Points in given space

Number of trials until first *success*

Multiple possible outcomes for each trial

Bernoulli distribution

- The bernoulli distribution is the “coin flip” distribution
- X is bernoulli if its probability function is:

$$X = \begin{cases} 1 & w.p. \quad p \\ 0 & w.p. \quad 1-p \end{cases}$$

  - X=1 is usually interpreted as a “success”
  - Examples

    X=1 for heads in coin toss

- Expectation:

$$E(X) = p(1) + (1-p)(0) = p$$

Bernoulli distribution

- Expectation:

$$E(X) = p(1) + (1-p)(0) = p$$

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= p(1)^2 + (1-p)(0)^2 - (p)^2 \\ &= p - p^2 = p(1-p) \end{aligned}$$

Binomial distribution

- The binomial distribution is just n independent bernoullis added up
- It is the number of "successes" in n trials
- If $Z_1, Z_2, \ldots, Z_n$ are bernoulli, then X is binomial:

$$X = Z_1 + Z_2 + \ldots + Z_n$$

- Testing for defects "with replacement"
  - Have many light bulbs
  - Pick one at random, test for defect, put it back
  - Pick one at random, test for defect, put it back
  - If there are many light bulbs, do not have to replace

Binomial distribution

- Let's figure out a binomial r.v.'s probability function
- Suppose we are looking at a binomial with $n=3$
- We want $P(X=0)$:
  - Can happen one way: 000
  - $(1-p)(1-p)(1-p)$
  - $(1-p)^3$
- We want $P(X=1)$:
  - Can happen three ways: 100, 010, 001
  - $p(1-p)(1-p)+(1-p)p(1-p)+(1-p)(1-p)p$
  - $3p(1-p)2$

Binomial distribution

- Let's figure out a binomial r.v.'s probability function
- Suppose we are looking at a binomial with n=3
- We want P(X=2):
  - Can happen three ways: 110, 011, 101
  - $pp(1-p)+(1-p)pp+p(1-p)p$
  - $3p^2(1-p)$
- We want P(X=3):
  - Can happen one way: 111
  - ppp
  - p^3

Binomial distribution

- Let’s figure out a binomial r.v.’s probability function
  - In general, for a binomial:

$$P_X(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

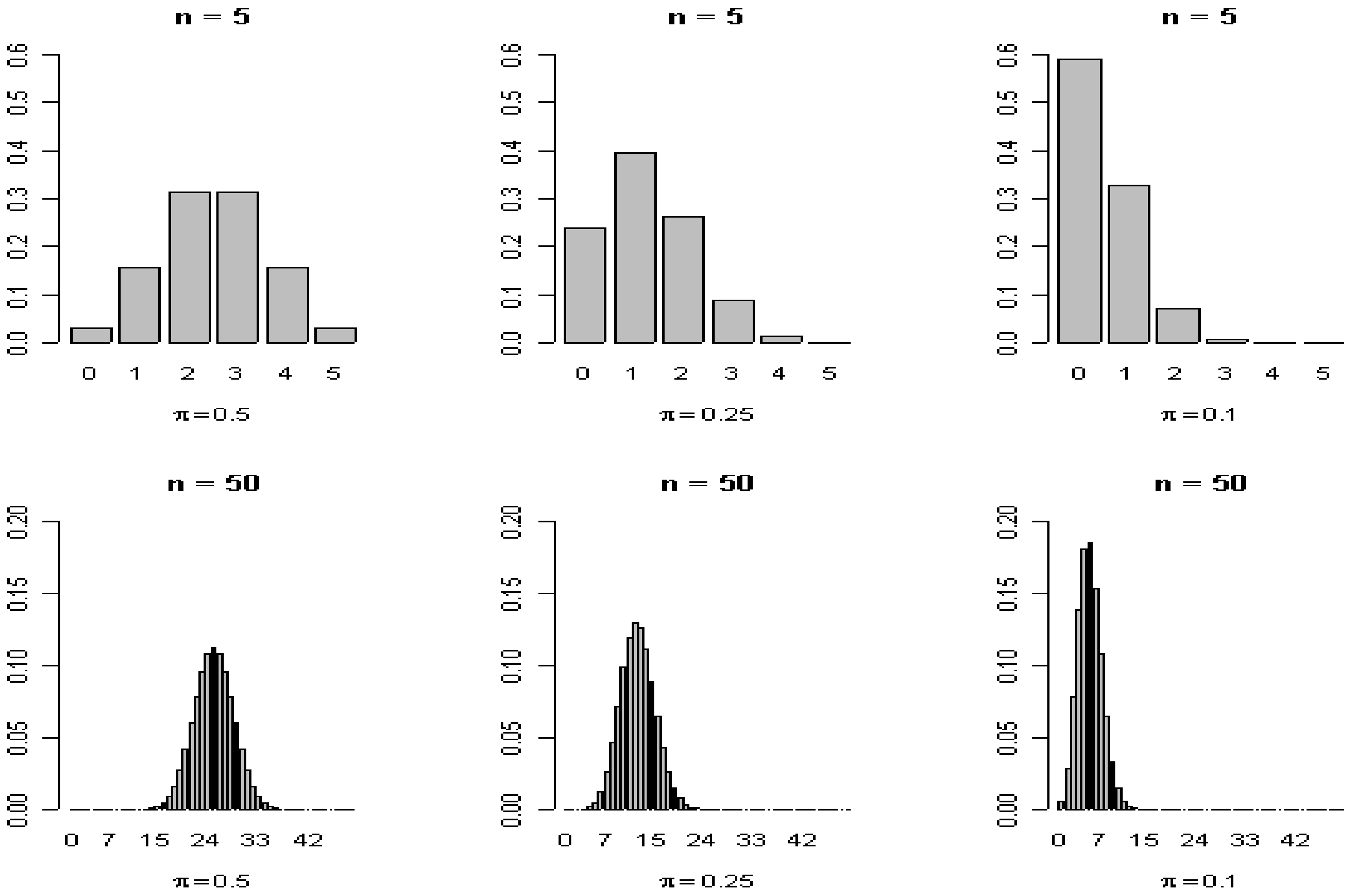
Binomial distribution

$$P_X(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Example n=5

		n	p	p	p	p
		5	0.5	0.25	0.1	0.05
x	x!	n!/(x!)(n-x)!	P(x)	P(x)	P(x)	P(x)
0	1	1	0.03125	0.2373	0.59049	0.7737809
1	1	5	0.15625	0.3955	0.32805	0.2036266
2	2	10	0.31250	0.2637	0.07290	0.0214344
3	6	10	0.31250	0.0879	0.00810	0.0011281
4	24	5	0.15625	0.0146	0.00045	0.0000297
5	120	1	0.03125	0.0010	0.00001	0.0000003
		sum =	1	1	1	1

Binomial Probability Density Function forms



As the n goes up, the distribution looks more symmetric and bell shaped.

Continuous Distributions

- **Normal Distribution**
- Log Normal Distribution
- Gamma Distribution
- **Chi Square Distribution**
- **F Distribution**
- **t Distribution**
- Weibull Distribution
- Extreme Value Distribution (Type I and II)

Foundations for much of statistical inference

Environmental variables

Time to failure, radioactivity

<u>Basis for statistical tests.</u>

Lifetime distributions

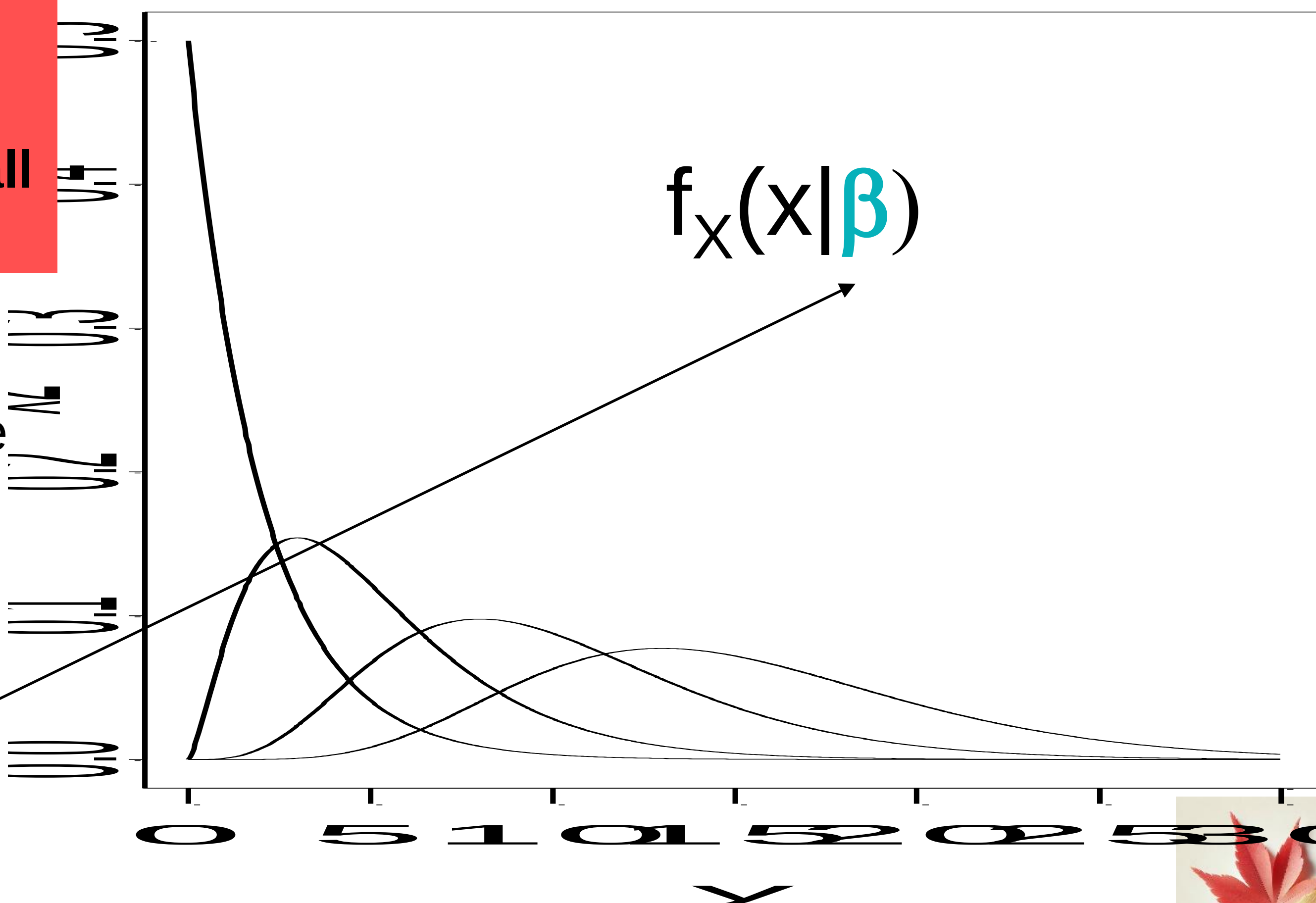
Continuous random variables are defined for continuous numbers on the real line. Probabilities have to be computed for all possible sets of numbers.

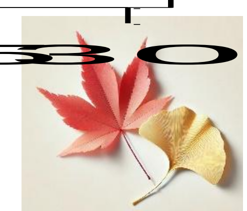
Probability Density Function

The pdf does not have to be symmetric, nor be defined for all real numbers.

The shape of the curve is determined by one or more **distribution parameters.**

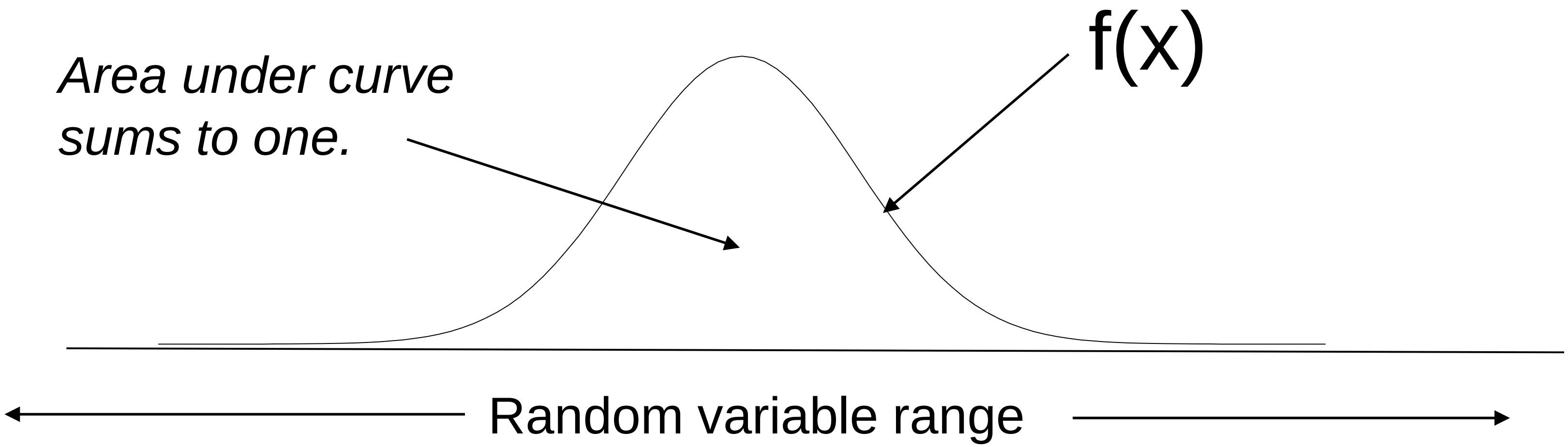
Chi Square density functions





Probability Density Function

A function which integrates to 1 over its range and from which event probabilities can be determined.



A theoretical shape - if we were able to sample the whole (infinite) population of possible values, this is what the associated histogram would look like.

A mathematical abstraction

Continuous Distribution Properties

Probability can be computed by integrating the density function.

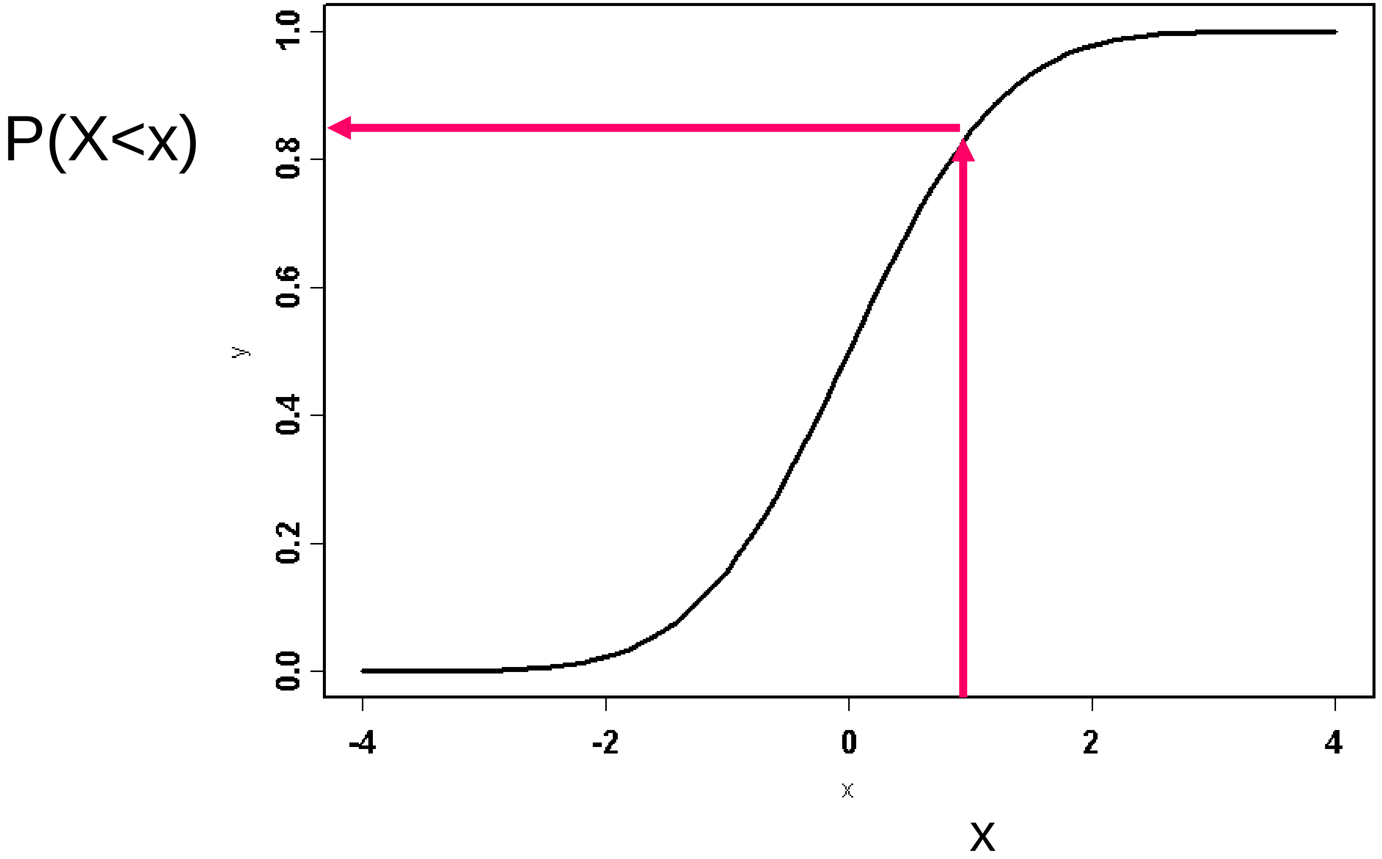
$$F(x_0) = P(X < x_0) = \int_{-\infty}^{x_o} f_X(x)dx$$

Continuous random variables only have positive probability for events which define intervals on the real line.

Any one point has **zero probability of occurrence**.

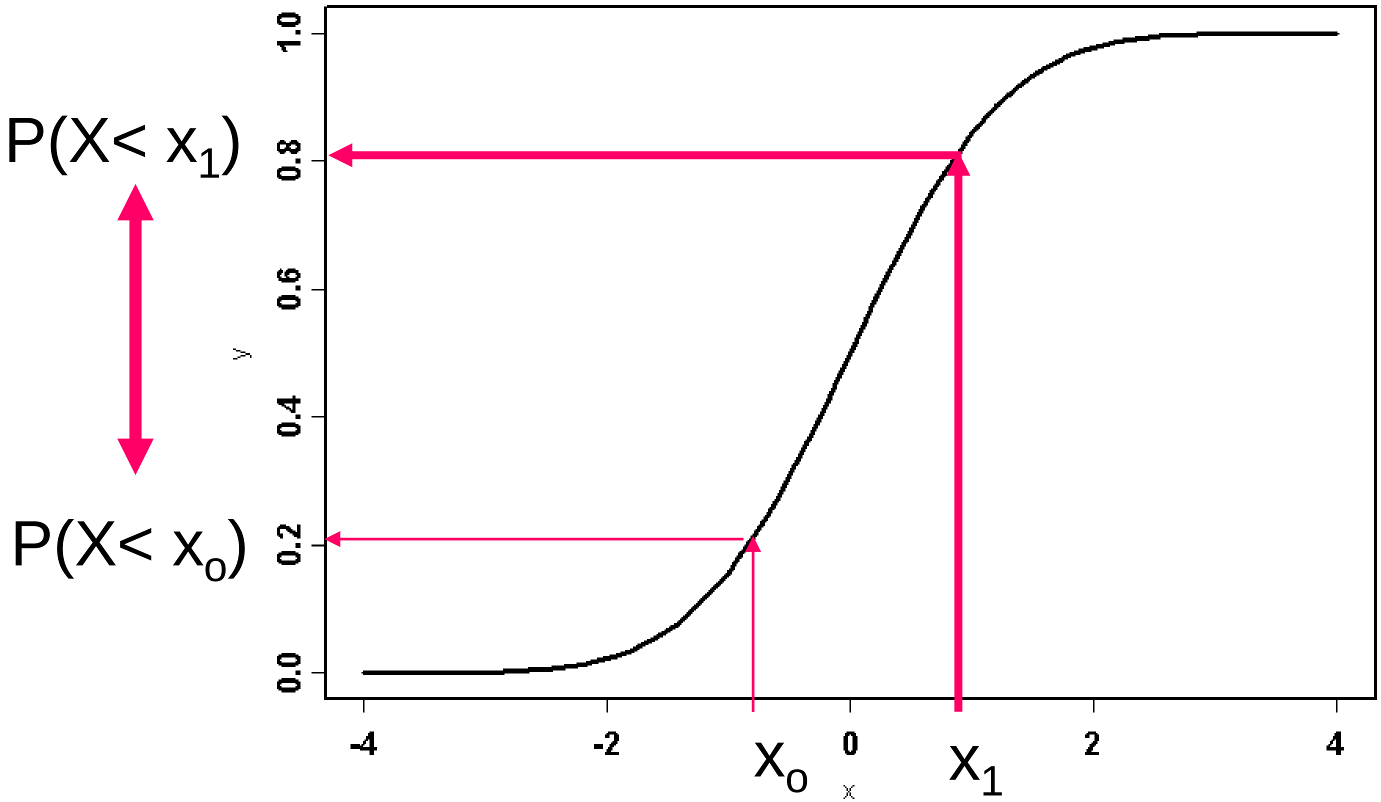
$$P(X = x_0) = \int_{x_0}^{x_0} f_X(x)dx = 0$$

Cumulative Distribution Function



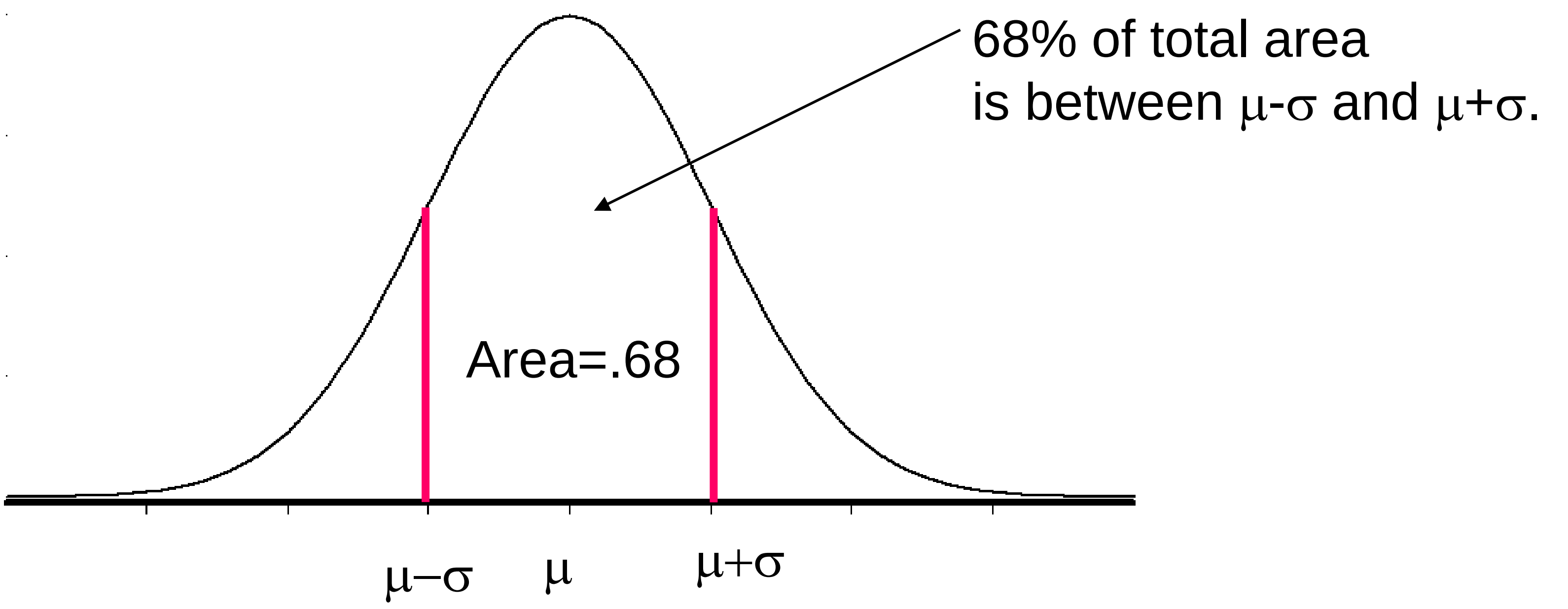
Using the Cumulative Distribution

$$P(x_o < X < x_1) = P(X < x_1) - P(X < x_o) = .8 - .2 = .6$$



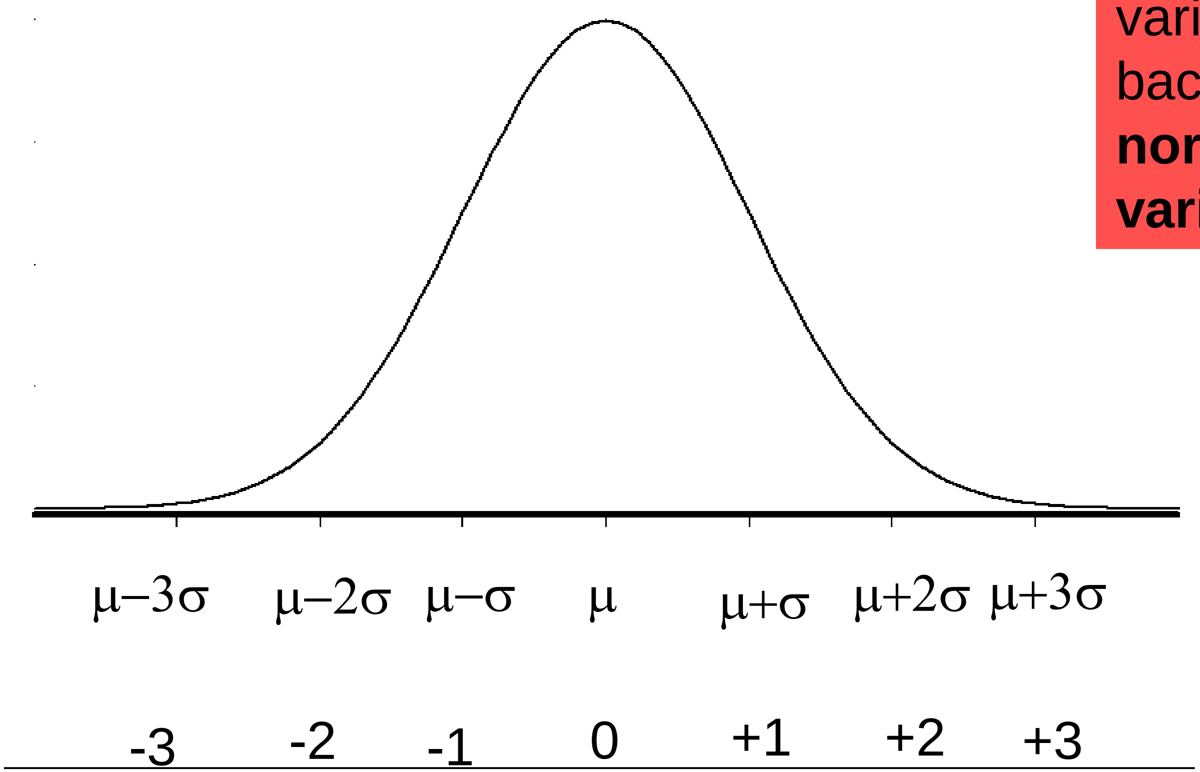
Normal Distribution

A symmetric distribution defined on the range -∞ to + ∞ whose shape is defined by two parameters, the **mean**, denoted μ, that centers the distribution, and the **standard deviation**, σ, that determines the spread of the distribution.



$$P(\mu - \sigma < X < \mu + \sigma) = .68$$

Standard Normal Distribution



All normal random variables can be related back to the **standard normal random variable**.

A Standard Normal random variable has mean 0 and standard deviation 1.

Estimation & Significance Testing

❖ Estimation (of population parameters)

Ex. "Based on GSS data, we're 95% confident that the population mean of the variable LONELY (no. of days in past week you felt lonely, $\bar{y}$ = 1.5, s = 2.2) falls between 1.4 and 1.6.

❖ Significance Testing

(Making decisions about hypotheses regarding "effects" and associations)
Ex. Article in Science 2008:
"We hypothesized that spending money on other people
has a more positive impact on happiness than spending money on oneself"

Statistical Inference : Estimation

Goal: How can we use sample data to estimate values of population parameters?

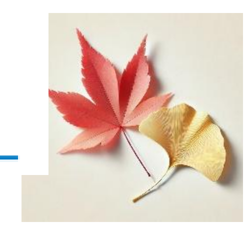
❖ **Point estimate**:
A single statistic value that is the “best guess” for the parameter value

❖ **Interval estimate**
An interval of numbers around the point estimate, that has a fixed “confidence level” of containing the parameter value.
Called a ***confidence interval***.
(Based on sampling distribution of the point estimate)

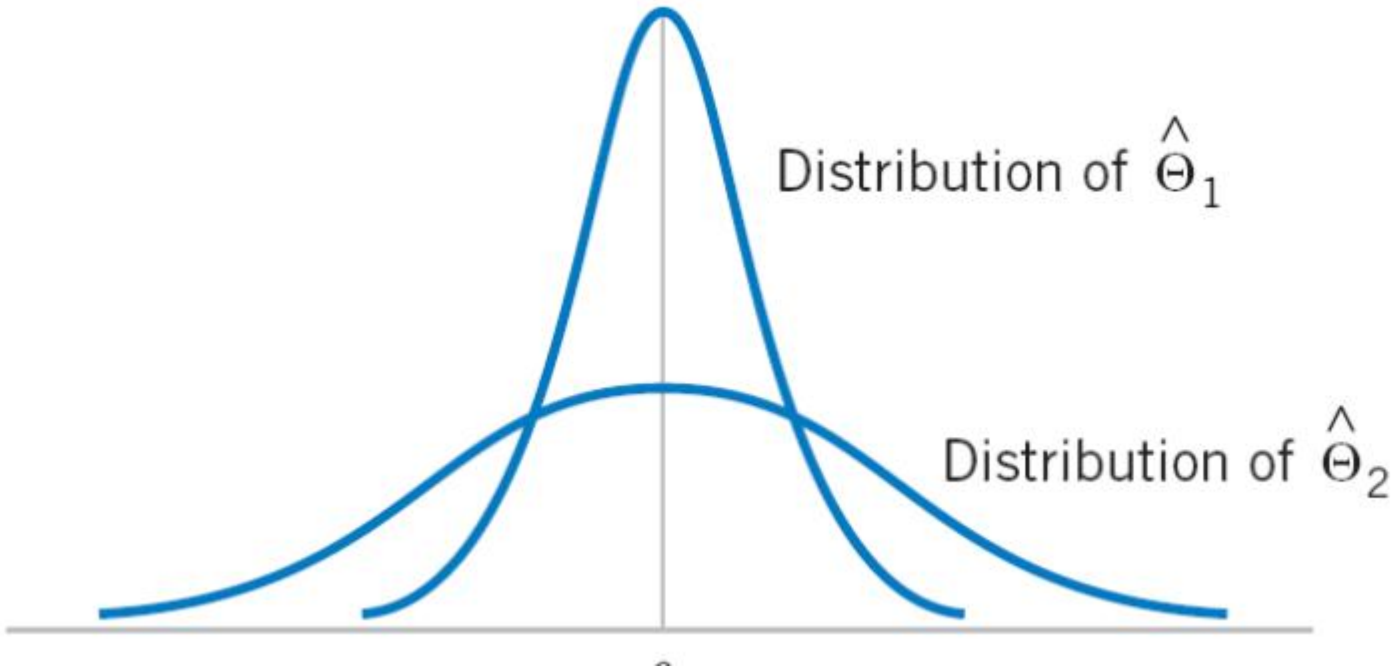
Point Estimation of Sampling

A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistics $\widehat{\Theta}$.

Unknown Parameter θ	Statistic $\hat{\Theta}$	Point Estimate $\hat{\theta}$
μ	$\bar{X} = \frac{\sum X_i}{n}$	$\bar{x}$
σ^2	$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$	s^2
p	$\hat{P} = \frac{X}{n}$	$\hat{p}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2 = \frac{\sum X_{1i}}{n_1} - \frac{\sum X_{2i}}{n_2}$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$



Point Estimation of Sampling



The sampling distributions of two unbiased estimators $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$.

Point Estimation of Sampling

❖ Good Estimator Condition: MVUE

If we consider all unbiased estimators of θ,the one with smallest variance is Called the **minimum variance unbiased estimator (MVUE)**.

The **mean square error** of an estimator $\widehat{\Theta}$ of the parameter θ is defined as

$$MSE\left(\widehat{\Theta}\right) = E(\widehat{\Theta} - \theta)^2$$

The **standard error** of a statistics is the standard deviation of its sampling distribution. If the standard error involves unknown parameters whose values can be estimated, substitution of these estimates into the standard error results in an **estimated standard error**.

Example 1 ~ internet usage data of 82 people

In [2]:
```
import pandas as pd
import numpy as np
data=pd.read_csv('./data/2.5.csv')

data.head(3)
```

Out[2]:

	value
0	22
1	22
2	20

In [3]:
```
data.value.describe()
```

Out[3]:
```
count    82.000000
mean     24.646341
std       4.089650
min      16.000000
25%      22.000000
50%      24.500000
75%      28.000000
max      33.000000
Name: value, dtype: float64
```

"bins" denotes the interval of data.

In [4]:
```
freq,bins=np.histogram(data, bins=6, range=(15.5,33.5))
bins
```

Out[4]:
```
array([15.5, 18.5, 21.5, 24.5, 27.5, 30.5, 33.5])
```


Example 1 ~ internet usage data of 82 people

In [5]:

```
freq_class=['15.5~18.5','18.5.~21.5','21.5~24.5','24.5~27.5','27.5~30.5','30.5~33.5']

freq_table=pd.DataFrame({'frequency':freq}, index=pd.Index(freq_class, name='class'))

freq_table
```

Out[5]:

class	frequency
15.5~18.5	7
18.5.~21.5	11
21.5~24.5	23
24.5~27.5	19
27.5~30.5	14
30.5~33.5	8

Example 1 ~ internet usage data of 82 people

In [6]:

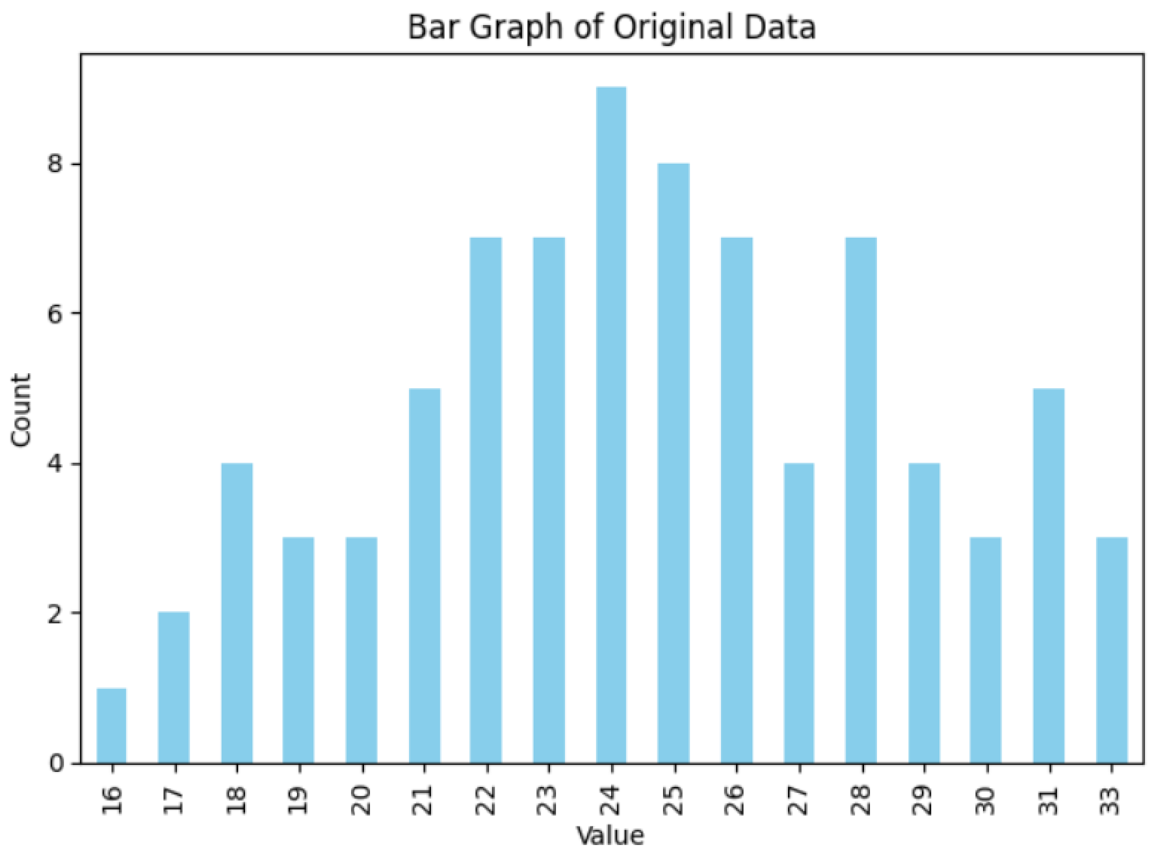
```
r_freq=freq/freq.sum()
cum_r_freq=np.cumsum(r_freq)
freq_table['relative frequency']=r_freq
freq_table['cumulative frequency']=cum_r_freq

freq_table
```

Out[6]:

class	frequency	relative frequency	cumulative frequency
15.5~18.5	7	0.085366	0.085366
18.5.~21.5	11	0.134146	0.219512
21.5~24.5	23	0.280488	0.500000
24.5~27.5	19	0.231707	0.731707
27.5~30.5	14	0.170732	0.902439
30.5~33.5	8	0.097561	1.000000

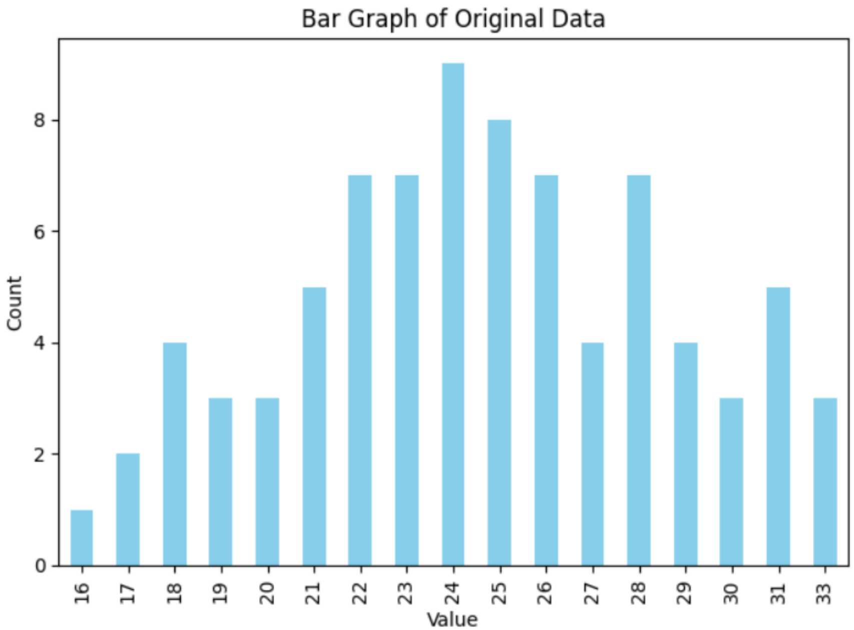
Example 1 ~ internet usage data of 82 people

```
# 'value' 열의 고유한 값별로 카운트
value_counts = data['value'].value_counts().sort_index()

# 막대그래프 그리기
value_counts.plot(kind='bar', color='skyblue')
plt.xlabel('Value')
plt.ylabel('Count')
plt.title('Bar Graph of Original Data')
plt.tight_layout()
plt.show()
```

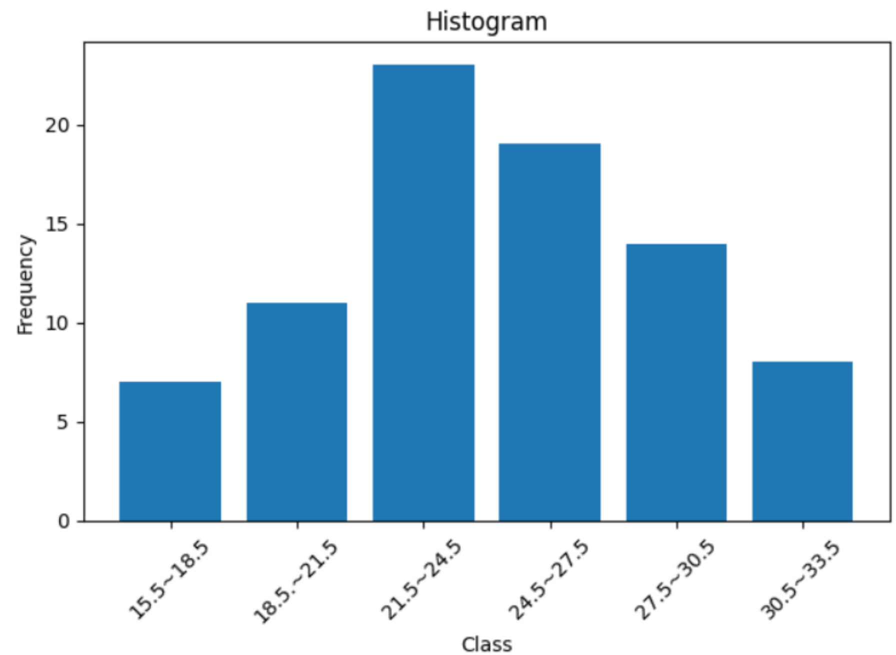


Example 1 ~ internet usage data of 82 people

population



sample



Example 1, Draw histogram for the sampled data according to the bin

Confidence Intervals

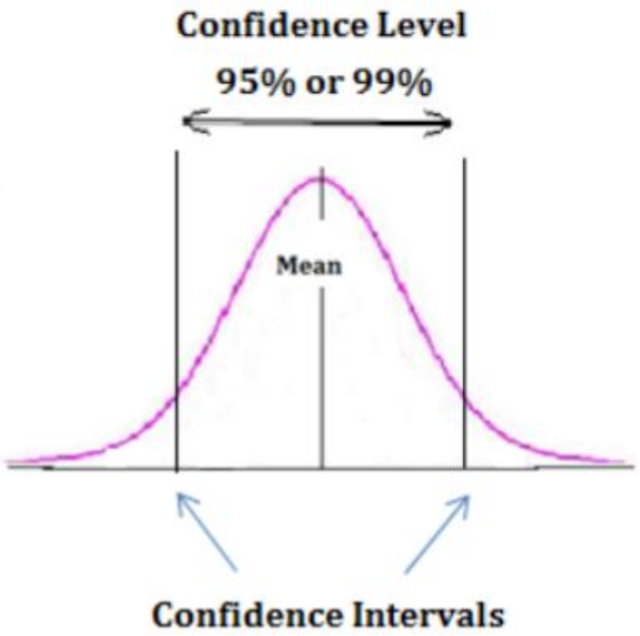
❖ A **confidence interval** (CI) is an interval of numbers believed to contain the parameter value.

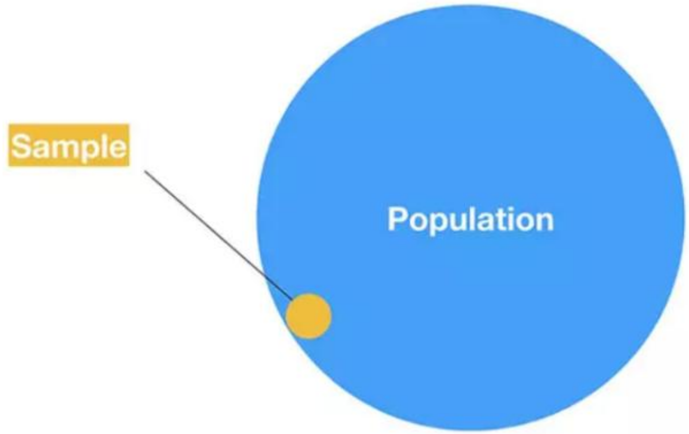
❖ The probability the method produces an interval that contains the parameter is called the **confidence level.**

Most studies use a confidence level close to 1, such as 0.95 or 0.99.

❖ Since studying the entire population is impossible, we estimate the range of parameters using sampled data.

→ The confidence interval measures how well the sampled data represents the population.





Confidence Intervals & Z-Score

Z-Score:

The z-score, also known as the standard score, measures how many standard deviations an element (or data point) is from the mean of the dataset.

The formula to calculate the z-score of a value x is given by:

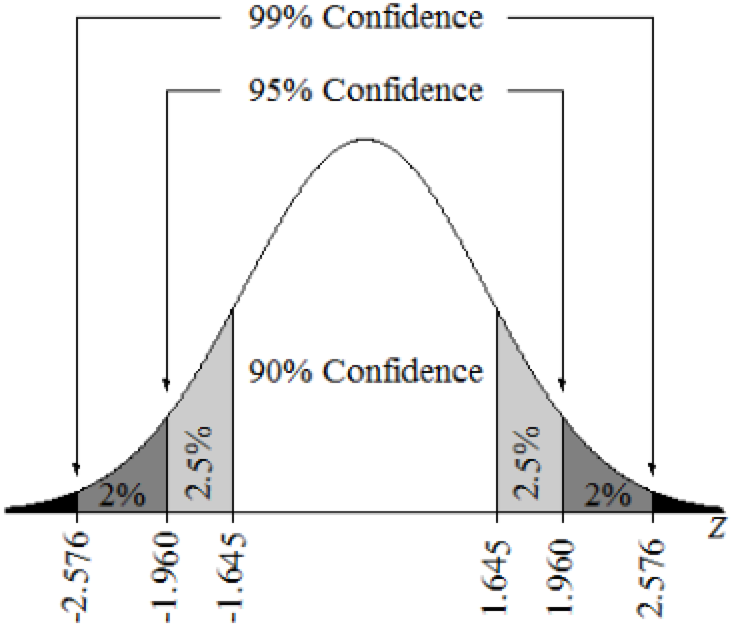
$$z = \frac{x-\mu}{\sigma}$$

x is the data point.

μ is the mean of the dataset.

σ is the standard deviation of the dataset.

Confidence Level	Z-Score
0.90	1.645
0.95	1.96
0.99	2.58



Confidence Intervals & Z-Score

Example:

observed sample : n=40

mean X= 175

Standard deviation s = 20

* Since we don't know σ(population standard deviation), observed standard deviation is used

$$CI = \bar{x} \pm Z \frac{s}{\sqrt{n}}$$

$\bar{x}$ is the mean of the observed data.

Z is the chosen value from the z-score table.

s is the standard deviation of the observed data.

n is the number of observations

In 95% CI,

$$175 \pm 1.960 \times \frac{20}{\sqrt{40}}$$

175cm ± 6.20cm

Thus, we estimate that the population's average lies within the confidence interval of 168.8cm to 181.2cm.

Confidence Interval for the mean

❖ In large random samples, the sample mean has approximately a normal sampling distribution with mean μ and standard error

$$\sigma_{\bar{y}} = \sigma / \sqrt{n}$$

Thus,

$$P(\mu - 1.96\sigma_{\bar{y}} \le \bar{y} \le \mu + 1.96\sigma_{\bar{y}}) = .95$$

❖ We can be 95% confident that the sample mean lies within 1.96 standard errors of the (unknown) population mean.

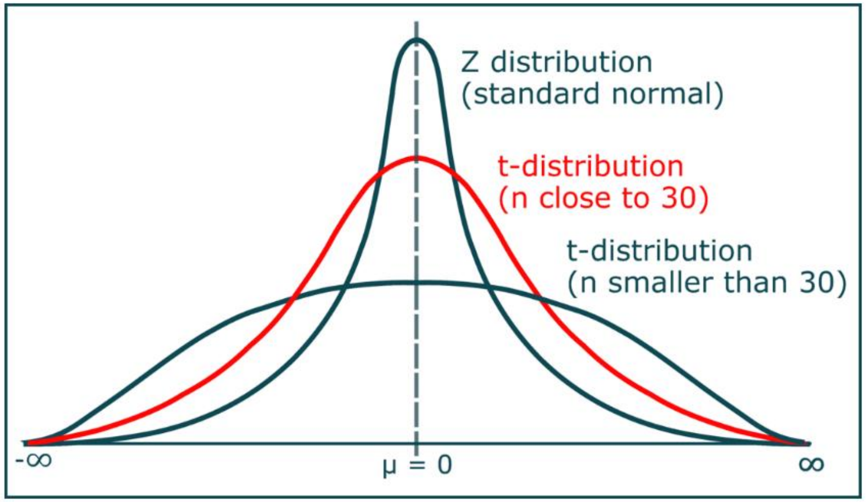
Confidence Interval for the mean

❖**Problem**: Standard error is **<u>unknown</u>** (σ is also a parameter). It is estimated by replacing σ with its point estimate from the sample data.

$$se = \frac{s}{\sqrt{n}}$$

95% confidence interval for μ :

$$\bar{y} \pm 1.96(se), \text{ which is } \bar{y} \pm 1.96\frac{s}{\sqrt{n}}$$



This works ok for "large *n*," because *s* then a good estimate of σ (and CLT applies). But **for small *n*(≤ 30)**, replacing σ by its estimate *s* introduces extra error, and CI is not quite wide enough unless we replace *z*-score by a slightly larger "***t*-score**."

t-score & Student t-distribution

❖ Student's t-distribution(t-distribution),

is a type of probability distribution that looks similar

to the normal distribution but generally has **heavier tails**.

It was introduced by William Sealy Gosset under the pseudonym "Student".

❖ Shape of t-distribution

- **Bell-shaped and symmetric:** Like the normal distribution,

  the t-distribution is bell-shaped and symmetric around its mean, which is 0.

- **Thicker tails:** Compared to the normal distribution, the t-distribution has thicker tails.

  This implies that the t-distribution gives more probability to values further from the mean

  As the sample size increases, the t-distribution approaches the normal distribution.

- **Degrees of Freedom (df):** The shape of the t-distribution is determined

  by its degrees of freedom, usually denoted as *df.*

t-score & Student t-distribution

❖ When to use the t-distribution:

- **Estimating the population in situations where the sample size is small (< 30).**

  This is especially relevant when the population standard deviation is unknown

  and you're using the sample standard deviation instead.

- **You're conducting a t-test.**

  The t-test is a statistical test that is used to determine

  if there's a significant difference between the means of two groups.

- **Population standard deviation is unknown:**

  Even with larger samples, if the population standard deviation is unknown,

  the t-distribution is the appropriate distribution to use,

  as we replace the population standard deviation

  with the sample standard deviation.

Example of T-table

❖ Suppose we compute a 95% confidence interval for the true systolic blood pressure using data in the subsample. Because the sample size is small, we must now use the confidence interval formula that involves t rather than Z.

$$CI = \bar{x} \pm \mathrm{T}\frac{s}{\sqrt{n}}$$

The sample size is n=10, the degrees of freedom (df) = n-1 = 9.

The t value for 95% confidence with df = 9 is t = 2.262.

T-table

Because the sample size is small, we need to use the t distribution.
For 95% confidence and df = n-1 = 9, t = 2,262.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318

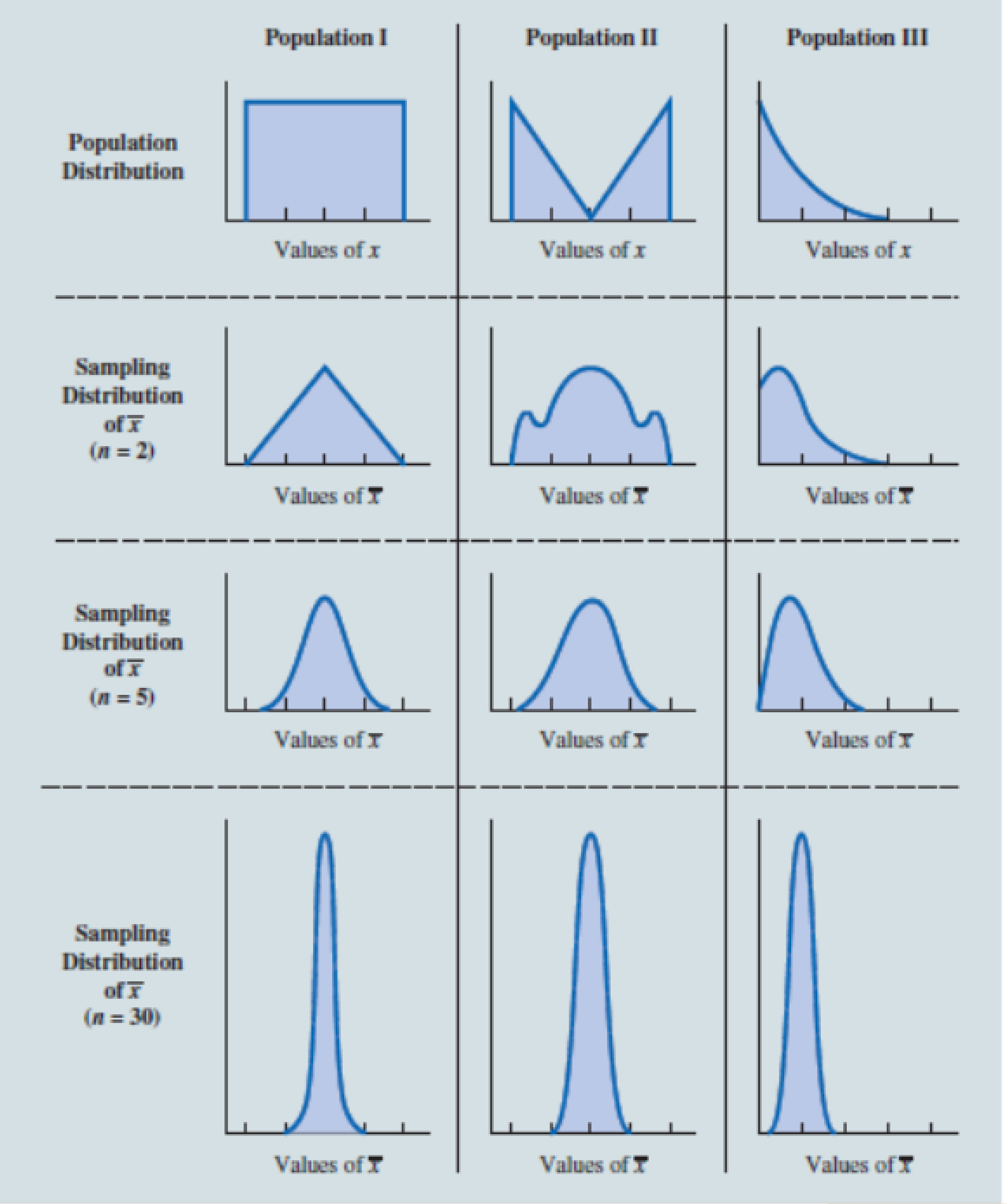
- Substituting the sample statistics and the t value for 95% confidence:

$$121.2 \pm 2.262 \frac{11.1}{\sqrt{10}} = 121.2 \pm 7.94 = \left(113.3,\ 129.1\right)$$

Based on this sample of size n=10, our best estimate of the true mean systolic blood pressure in the population is 121.2. Based on this sample, we are 95% confident that the true systolic blood pressure in the population is between 113.3 and 129.1. Note that the margin of error is larger here primarily due to the small sample size.

Illustration of the central limit theorem

- The Central Limit Theorem states that the distribution of sample means approaches a normal distribution, regardless of the population's original distribution, as the sample size becomes large enough.



Statistical Hypothesis H_0 vs. H_1

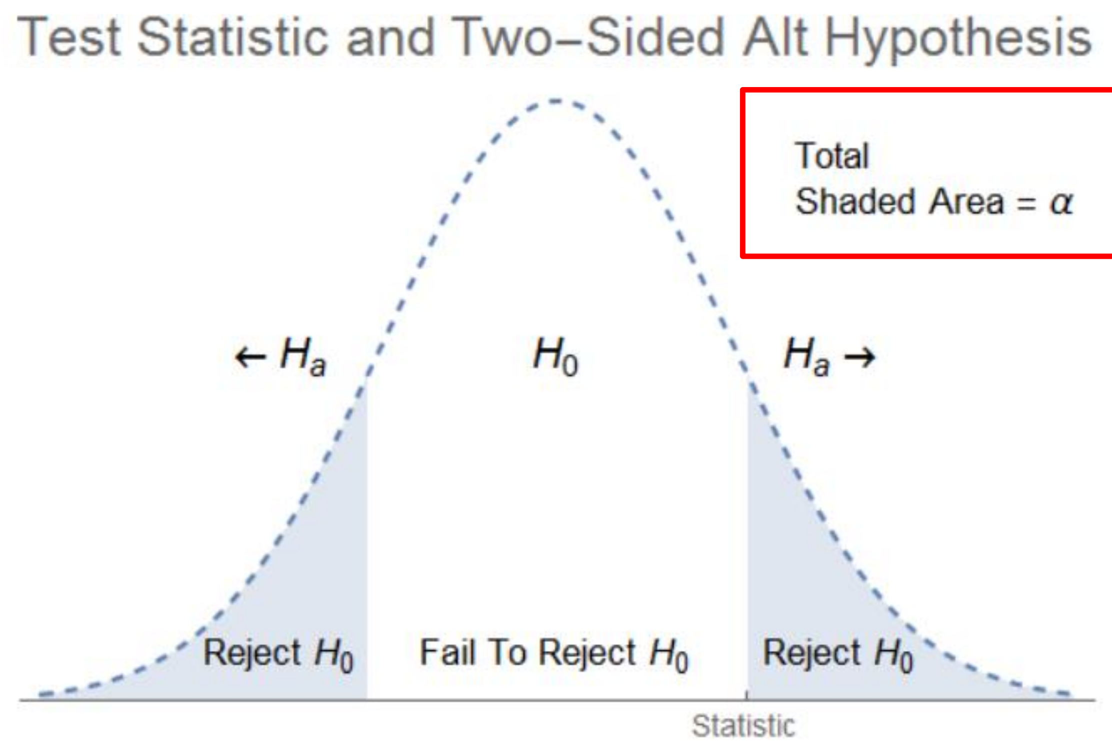
hypothesis	Ho	H1
ideal case	rejected	Accepted

❖ Key Statistical Concepts:

- **Null Hypothesis (Ho):**
  Represents the idea that there's no significant effect or difference.
  Examples include: "The research results are not meaningful,"
  "There's no difference," or "The product has no effect."
  When data, under the assumption that Ho is true,
  shows very different results from the null hypothesis,
  the null hypothesis is rejected, indicating statistical significance.

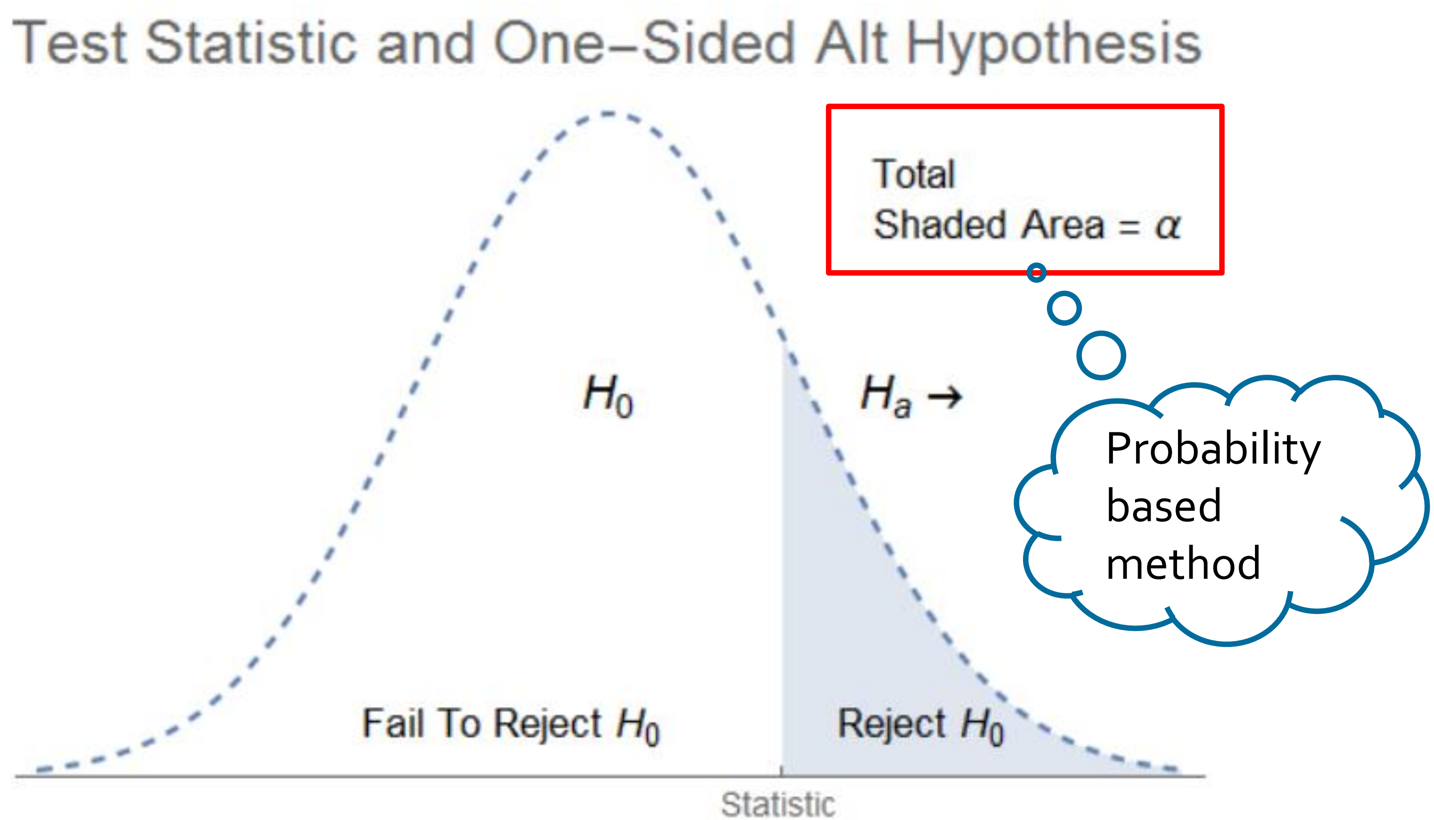
- **Alternative Hypothesis (H1 or Ha):**
  Represents the opposite idea of the null hypothesis.
  Examples include: "The research results are meaningful," "There is a difference,"
  or "The product has an effect."
  When Ho is rejected, the alternative hypothesis H1 is accepted,
  suggesting that the results are statistically significant.

Testing statistical Hypothesis H_0 (Type-1)



If **p <** **α**, then we view **the data as sufficiently unlikely to have occurred by chance**: We reject the null hypothesis in favor of the alternative hypothesis and say that **the evidence against the null hypothesis is statistically significant**.

Testing statistical Hypothesis H_0 (Type-1)



If **p < <u>α</u>**, then we view **the data as sufficiently unlikely to have occurred by chance**:
We reject the null hypothesis in favor of the alternative hypothesis and say that
the evidence against the null hypothesis is statistically significant.

Testing statistical Hypothesis H_0 (Type-1)

The power of a statistical test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis is true.

- The power is computed as $\boxed{\mathbf{1 - \beta}}$, and power can be interpreted as

the probability of correctly rejecting a false null hypothesis.

We often compare statistical tests by comparing their **power** properties.

•For example, consider the propellant burning rate problem when
we are testing H_0 : μ = 50 cm per second against H_1 : μ not equal 50 cm per second .
Suppose that the true value of the mean is μ = 52.
When $n = 10$, we found that β = 0.2643,
so the power of this test is 1 - β = 1 - 0.2643 = 0.7357 when μ = 52.

Example of Testing Hypothesis 1

❖ Hypothesis Testing for Customer Waiting Time at a Counter:

- **Background:**
  - The known average waiting time for customers at a certain counter is 11 minutes.
  - It's assumed that the waiting time for customers during a specific time slot follows a normal distribution.
  - Sampled waiting times are: 8, 10, 10, 7, 9, 12, 10, 8, 7, 9 minutes.

- **Hypotheses:**
  - **Null Hypothesis H_0:** The average waiting time μ = 11.
  - **Alternative Hypothesis H_1 :** The average waiting time $\mu \neq 11$.

- **Sample Statistics:**
  - Sample mean $\bar{x}$ = 9 minutes.
  - Sample variance s^2 = 2.44 (calculated from given data as 22/9).

Example of Testing Hypothesis

❖ Hypothesis Testing for Customer Waiting Time at a Counter:

- **Test Statistic and Rejection Region:**
  Using an alpha level of 0.05 and a t-distribution (because of the small sample size) with 9 degrees of freedom:
    Critical t-value $t_{0.025}(9) = 2.26$.
    Refer to p.47 T-table
    The rejection region for H_0 is when $\bar{x} \leq 9.88$ or $\bar{x} > 12.12$.

- **Results:**
  The observed sample mean $\bar{x} = 9$ falls into the rejection region, so the null hypothesis H_0 is rejected.
  This suggests that the average waiting time during that specific time slot is not 11 minutes.

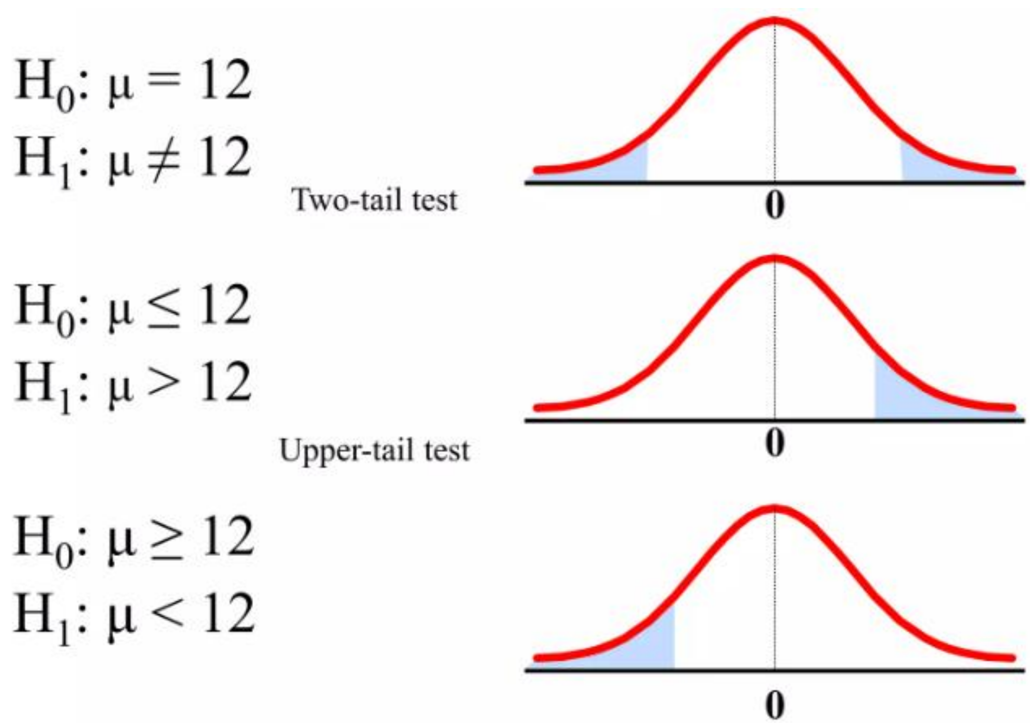
P-value approach

❖ **p-value** :

significant(critical) probability (of Ho Hypothesis) calculated from sample.

→ The p-value can be understood as the area in the tail region from the actual computed test statistic value.

The blue area = p-value



Practical Comments on Hypothesis Testing

The Seven-Step Procedure

Only three steps are really required:

1. Specify the hypothesis (two-, upper-, or lower-tailed).
2. Specify the test statistic to be used (such as z_0).
3. Specify the criteria for rejection (typically, the value of α, or the P-value at which rejection should occur).

Practical Comments on Hypothesis Testing

Statistical versus Practical Significance

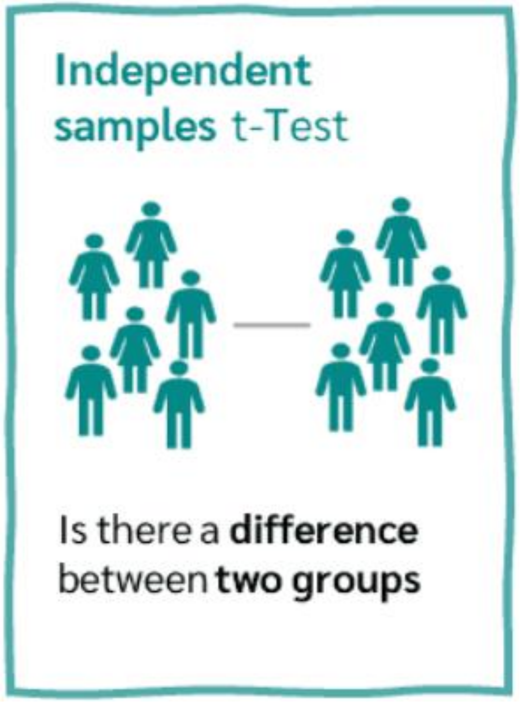
Sample Size n	P-Value When $\overline{x} = 50.5$	Power (at $\alpha = 0.05$) When $\mu = 50.5$
10	0.4295	0.1241
25	0.2113	0.2396
50	0.0767	0.4239
100	0.0124	0.7054
400	5.73×10^{-7}	0.9988
1000	2.57×10^{-15}	1.0000

Hypothetical Testing Example 2
Test for differences between two populations(samples)

Example:

We would like to find out whether there is a difference between chemical analysis and X-ray analysis methods in estimating iron content. Five specimens were split in two, chemical analysis was used on one piece, and iron content was measured on the other piece using X-ray analysis.

The results are as follows.



	0	1	2	3	4	5	6	7	8	9
g	A	A	A	A	A	B	B	B	B	B
x	2.0	2.0	2.3	2.1	2.4	2.2	1.9	2.5	2.3	2.4

The sample mean and standard deviation of the chemical analysis method and the sample mean and standard deviation of the X-ray analysis method are obtained as follows

g	count	mean	std	min	25%	50%	75%	max
A	5.0	2.16	0.181659	2.0	2.0	2.1	2.3	2.4
B	5.0	2.26	0.230217	1.9	2.2	2.3	2.4	2.5

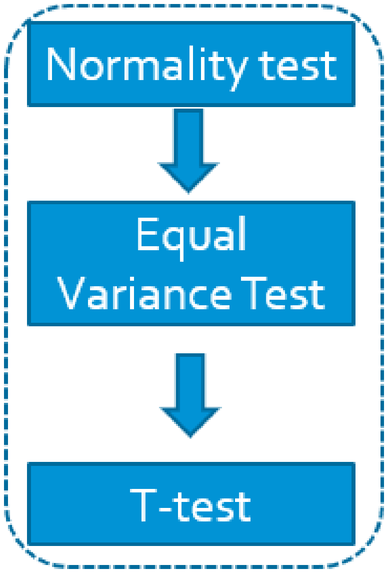
Test for differences between two populations

```
import numpy as np
import pandas as pd

#x:iron content  g: Analysis method
x=np.array([2.0,2.0,2.3,2.1,2.4,2.2,1.9,2.5,2.3,2.4])
g=np.repeat(np.array(['A', 'B']), 5)
d={'g':g, 'x':x}
data=pd.DataFrame(data=d)
data.T
```

g	A	A	A	A	A	B	B	B	B	B
x	2.0	2.0	2.3	2.1	2.4	2.2	1.9	2.5	2.3	2.4

```
#group  A:chemical method  B:X ray
A=data[data.g=='A']
B=data[data.g=='B']

#statistical analysis according to the group
data.groupby("g").x.describe()
```

	count	mean	std	min	25%	50%	75%	max
g								
A	5.0	2.16	0.181659	2.0	2.0	2.1	2.3	2.4
B	5.0	2.26	0.230217	1.9	2.2	2.3	2.4	2.5



groupby('g') denotes that describe iron content according to the analysis method

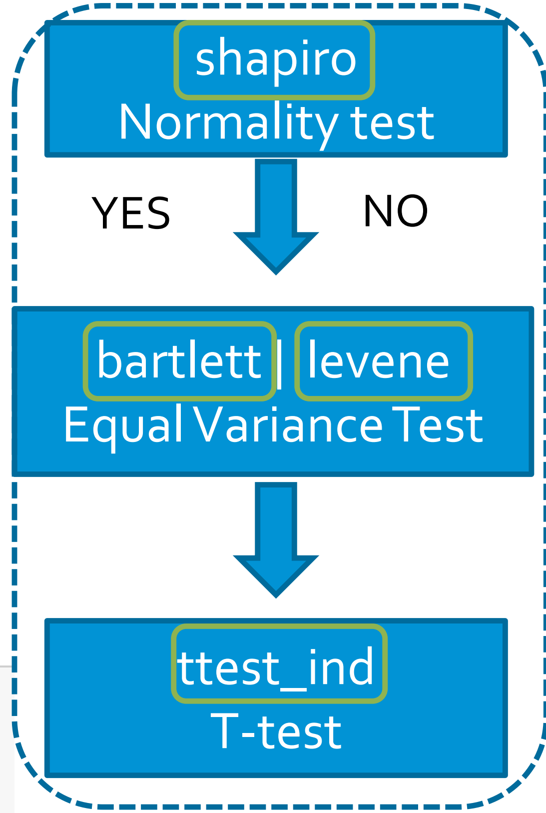
Test for differences between two populations

	0	1	2	3	4	5	6	7	8	9
g	A	A	A	A	A	B	B	B	B	B
x	2.0	2.0	2.3	2.1	2.4	2.2	1.9	2.5	2.3	2.4

❖ The ***normality** and **homoscedasticity** of the data* must be determined in order to perform a ***t-test*** for the difference between the two groups. Most programs use the **shapiro** function, and when the size of the data is very large, the **Anderson-Darling test** is sometimes used.

```
#normality testing
from scipy.stats import shapiro
shapiro(x)

```

```
ShapiroResult(statistic=0.9437551498413086, pvalue=0.5955022573471069)
```

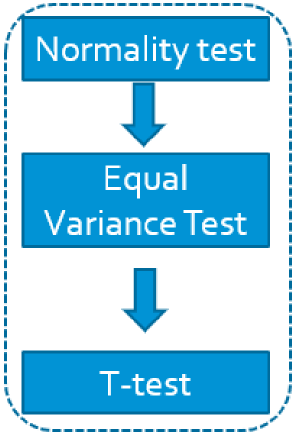


If the p value for the data test is greater than 0.05,
Ho is adopted and the conclusion is drawn that 'the data follows a normal distribution.'
If the p value is less than 0.05,
the conclusion is that 'the data do not follow a normal distribution.'

Test for differences between two populations

```
#normality testing
from scipy.stats import shapiro
shapiro(x)

```

ShapiroResult(statistic=0.9437551498413086, pvalue=0.5955022573471069)



In [4]:

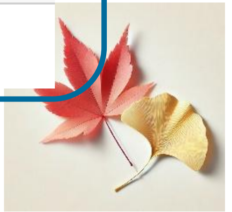
```
#Equal variance test - Bartlett test (when normality is satisfied)
from scipy import stats
stats.bartlett(A.x, B.x)

```

Out[4]: BartlettResult(statistic=0.19769157819919453, pvalue=0.6565906251784377)

In [5]:

```
#Equal variance test - Levene test (when normality is not satisfied)
from scipy import stats
stats.levene(A.x, B.x)

```

Out[5]: LeveneResult(statistic=0.055555555555555569, pvalue=0.8195856784525775)



Test for differences between two populations

In [5]:
```python
#Equal variance test - Levene test (when normality is not satisfied)
from scipy import stats
stats.levene(A.x, B.x)

```
Out[5]: LeveneResult(statistic=0.05555555555555569, pvalue=0.8195856784525775)

In [6]:
```python
#T test - two-sided test, equal variance assumption
from scipy.stats import ttest_ind
ttest_ind(A.x, B.x, equal_var=True)

```
Out[6]: Ttest_indResult(statistic=-0.7624928516630208, pvalue=0.4676497723369858)

P-value
→ t-test
formula

Normality test
↓
Equal Variance Test
↓
T-test

The t value calculated in the program is -0.7624 and
the p value is 0.4676, which is greater than 0.05.
Therefore, the assumption
"Ho: the iron content estimated according to chemical analysis
and X-ray analysis is the same" is adopted.

Hypothetical Testing Example 3

Test for paired data

The problem described *<u>previous compares two independent populations</u>*.
This case is different in that
<u>it compares two data groups from one population</u>.

❖ This is called **<u>a test for paired data</u>**.

❖ **Example:**

Various food additives are used in processed foods, and sorbic acid is typically used as a preservative for long-term preservation. The Department of Food and Nutrition at the University of Virginia investigated the residual amount of sorbic acid (unit ppm/ham) before and after storage of processed ham.
The residual amount of soribic acid in ham before storage was measured,
and the remaining amount of soribic acid was measured after 60 days of storage.
The results were as follows.

	0	1	2	3	4	5	6	7
y	116	96	239	329	437	597	689	576
x	224	270	400	444	590	660	1400	680

Hypothetical Testing Example 3

Test for paired data

In [5]:

```
import numpy as np
import pandas as pd

x=np.array([224,270,400,444,590,660,1400,680])
y=np.array([116,96,239,329,437,597,689,576])
d={'y':y, 'x':x}
data=pd.DataFrame(data=d)
data.T
```

Out[5]:

	0	1	2	3	4	5	6	7
y	116	96	239	329	437	597	689	576
x	224	270	400	444	590	660	1400	680

In [2]:

```
from scipy.stats import ttest_rel
#Paired T test - two-tailed test
ttest_rel(x,y)

```

Out[2]:

```
Ttest_relResult(statistic=2.673117820270042, pvalue=0.031855388760108426)
```

In a t-distribution with 7 degrees of freedom, the p value of the two-sided test is 0.0319.
Therefore, the p value for whether the residual amount of sorbic acid decreased was 0.0159.
Since it is less than the significance level of 0.05,
Ho can be rejected.
Therefore, the residual amount of sorbic acid in stored ham decreased compared to before storage.

Hypothetical Testing Example 4
Test for one sample

❖ **Example:**

Let us assume that the number of days of service (X) for a new employee in a financial industry follows a normal distribution.
If the number of days of service for the 9 employees who left the company is as follows, can it be claimed that the number of days of service for this new employee in the financial industry is 1950 days?

Ho="The number of days of service for a new employee in the financial industry is 1,950 days."

In [1]:

```python
import numpy as np
x=np.array([2000, 1975, 1900, 2000, 1950, 1850, 1950, 2100, 1975])

# t-test for one sample
import scipy.stats as stats
stats.ttest_1samp(x, popmean=1950)
```

Out[1]:

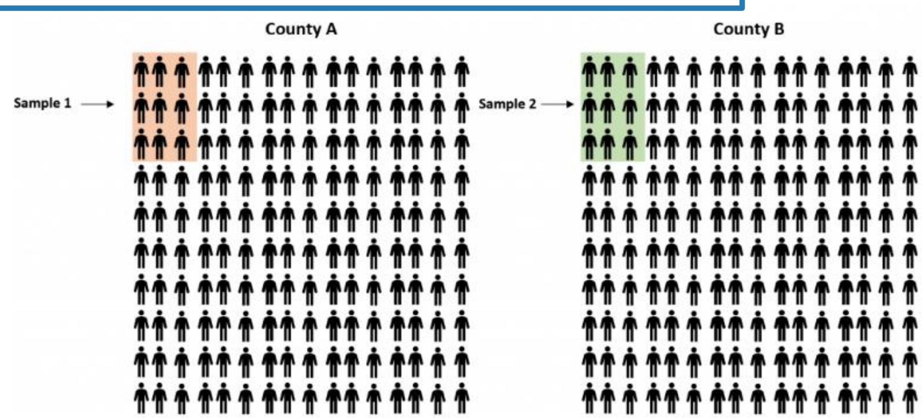
```
Ttest_1sampResult(statistic=0.7184212081071027, pvalue=0.4929429157541276З)
```

P-value is greater than 0.05. thus, we cannot reject Ho.
Thus, we can say that
The number of days of service for a new employee in the financial industry is 1,950 days."

Hypothetical Testing Example 5
Test for difference between two population proportions

❖ **Example:**
We want to find out whether there is a difference in the market share of
a company's products between regions A and B.
As a result of surveying 80 people each in Region A and Region B,
56 people in Region A and 44 people in Region B are using this company's products.
Can it be said that Region A's market share is higher than Region B?
Since we want to find out whether the market share (Pa) of region A
is higher than the market share (Pb) of region B,
the null hypothesis and alternative hypothesis are as follows.

$$\boxed{\begin{array}{l} H_0: p_A \leq p_B \rightleftharpoons p_A - p_B \leq 0 \\ H_1: p_A > p_B \rightleftharpoons p_A - p_B > 0 \end{array}}$$



Hypothetical Testing Example 5
Test for difference between two population proportions

In [1]:
```
import pandas as pd

data=pd.DataFrame([[56,24],[44,37]], index=['A','B'], columns=['use', 'unuse'])
data

```

Out[1]:

	use	unuse
A	56	24
B	44	37

In [3]:
```
#Population ratio test
from scipy.stats import fisher_exact
fisher_exact(data, alternative='greater')

```

Out[3]:
```
(1.9621212121212122, 0.029289557246881863)
```

The population ratio test in the program uses Fisher's exact test function. Alternative is defined as less, greater, or two-sided depending on the hypothesis.

P-value is less than 0.5. thus, we can reject Ho.
Thus, we can say that
“Region A's market share is higher than Region B”.

Hypothetical Testing Example 6
Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank Test is the non-parametric version of the paired samples t-test. It is used to test whether or not there is a significant difference between two population means when the distribution of the differences between the two samples cannot be assumed to be normal.

Example:

Researchers want to know if a new fuel treatment leads to a change in the average mpg of a certain car. To test this, they measure the mpg of 12 cars with and without the fuel treatment.

Use the following steps to perform a Wilcoxon Signed-Rank Test to determine if there is a difference in the mean mpg between the two groups.

Step 1: Create the data.

First, we'll create two arrays to hold the mpg values for each group of cars:

```
group1 = [20, 23, 21, 25, 18, 17, 18, 24, 20, 24, 23, 19]
group2 = [24, 25, 21, 22, 23, 18, 17, 28, 24, 27, 21, 23]
```


Hypothetical Testing Example 6
Wilcoxon Signed-Rank Test

First, we'll create two arrays to hold the mpg values for each group of cars:

```
group1 = [20, 23, 21, 25, 18, 17, 18, 24, 20, 24, 23, 19]
group2 = [24, 25, 21, 22, 23, 18, 17, 28, 24, 27, 21, 23]
```

Step 2: Conduct a Wilcoxon Signed-Rank Test.

Next, we'll use the wilcoxon() function from the scipy.stats library to conduct a Wilcoxon Signed-Rank Test, which uses the following syntax:

wilcoxon(x, y, alternative='two-sided')

where:

- **x:** an array of sample observations from group 1
- **y:** an array of sample observations from group 2
- **alternative:** defines the alternative hypothesis. Default is 'two-sided' but other options include 'less' and 'greater.'

Hypothetical Testing Example 6
Wilcoxon Signed-Rank Test

```
import scipy.stats as stats

#perform the Wilcoxon-Signed Rank Test
stats.wilcoxon(group1, group2)

(statistic=10.5, pvalue=0.044)
```

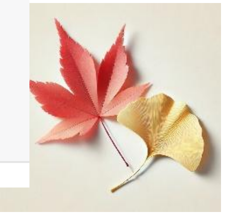
The test statistic is **10.5** and the corresponding two-sided p-value is **0.044**.

Step3: Interpret the results
In this example, the Wilcoxon Singed Rank Test uses the following null and alternative hypotheses:
H0: The mpg is equal between two groups
H1: The mpg is not equal between two groups

Since the p-value(0.044) is less than 0.05, we reject null hypothesis.
We have sufficient evidence that
the true mean mpg is not equal between the two groups.

Paired Sampled T-Test vs. Wilcoxon Signed-Rank Test

1. **Assumptions**:
   - **Paired Sampled t-test**: This test assumes that the data is normally distributed. Additionally, considerations about equal variances (the assumption that the two groups have the same variance) might be necessary.
   - **Wilcoxon Signed Rank Test**: This is a non-parametric method, which means it doesn't require the data to follow a specific distribution (e.g., normal distribution).
2. **Type of Data**:
   - **Paired Sampled t-test**: Used for continuous data.
   - **Wilcoxon Signed Rank Test**: Can be used for ordinal (ranked or ordered) data, as well as continuous data.
3. **Calculation Approach**:
   - **Paired Sampled t-test**: Based on the difference in means.
   - **Wilcoxon Signed Rank Test**: Based on the ranks of differences. It calculates the differences between the two related samples, then ranks these differences by their absolute values, and assigns signs to the ranks based on the original sign of the difference.
4. **Robustness**:
   - **Paired Sampled t-test**: Sensitive to outliers and violations of the normality assumption.
   - **Wilcoxon Signed Rank Test**: More robust to outliers.



Hypothetical Testing Example 7
Chi square Test

The **Chi-squared test** assesses associations between categorical variables.
It compares observed frequencies to expected frequencies under the assumption of independence. A significant result suggests a potential relationship between variables.
Commonly used in *contingency tables*,
it requires sufficient sample size and expected frequencies to ensure validity.

Example:
A certain card company believes there might be a relationship between a customer's grade (A, B, C, D: with A being the highest grade) and the amount they spend using the card.
To test for independence, they obtained the following contingency table.
They surveyed the spending amounts of 860 customers and categorized the spending into five levels and the customer grades into four levels.

level	A	B	C	D
amount				
10under	21	42	60	5
10~20	15	122	45	14
20~40	94	100	16	30
40~70	120	65	20	18
70upper	32	9	12	20

Hypothetical Testing Example 7

Chi square Test

level	A	B	C	D
amount				
10under	21	42	60	5
10~20	15	122	45	14
20~40	94	100	16	30
40~70	120	65	20	18
70upper	32	9	12	20

In [18]:

```python
import pandas as pd
data=pd.read_csv('ex7-4.csv')
data.head()
```

Out[18]:

	amount	level	count
0	10under	A	21
1	10~20	A	15
2	20~40	A	94
3	40~70	A	120
4	70upper	A	32

In [14]:

```python
#construct frequency table
pd.crosstab(index=data['amount'], columns=data['level'], values=data['count'], aggfunc=sum,₩
            margins=True, margins_name='전체')

```

Out[14]:

level	A	B	C	D	전체
amount					
10under	21	42	60	5	128
10~20	15	122	45	14	196
20~40	94	100	16	30	240
40~70	120	65	20	18	223
70upper	32	9	12	20	73
전체	282	338	153	87	860

"When constructing a frequency table using the pd.crosstab function, if margins is set to True, it calculates the row and column totals. margins_name defines the name for the row and column of these totals(전체).

Hypothetical Testing Example 7

Chi square Test

In [13]:

```python
#construct probability table
pd.crosstab(index=data['amount'], columns=data['level'], values=data['count'], aggfunc=sum, ₩
            margins=True, margins_name='전체', normalize='index').round(4)

```

Out[13]:

level amount	A	B	C	D
10under	0.1641	0.3281	0.4688	0.0391
10~20	0.0765	0.6224	0.2296	0.0714
20~40	0.3917	0.4167	0.0667	0.1250
40~70	0.5381	0.2915	0.0897	0.0807
70upper	0.4384	0.1233	0.1644	0.2740
전체	0.3279	0.3930	0.1779	0.1012

When constructing a probability table, one uses the **normalize** parameter in pd.crosstab. If normalize is set to 'all', it displays the overall percentage; if set to 'index', it shows the row percentage; and if set to 'columns', it presents the column percentage.

In [17]:

```python
#Chi-squared test and  constructing Expected Frequency Table
from scipy.stats import chi2_contingency
d_table=pd.crosstab(index=data['amount'], columns=data['level'], values=data['count'], aggfunc=sum, ₩
                    margins=True, margins_name='전체')

chi,p,df,expected=chi2_contingency(d_table)

expected

```

Out[17]:

```
array([[ 41.97209302,  50.30697674,  22.77209302,  12.94883721,
        128.        ],
       [ 64.26976744,  77.03255814,  34.86976744,  19.82790698,
```

Hypothetical Testing Example 7
Chi square Test

In [16]:

```python

expected_table=pd.DataFrame(data=expected, index=d_table.index, columns=d_table.columns)
expected_table

print(chi, p)
```

```
252.05782411526025 4.392425562427717e-42
```

The result of the Chi-squared test indicates a Chi-squared test statistic value of 252.0578, and the p-value is very small, leading to the rejection of H_0.
Therefore, card spending amount and customer grade are not independent.
In other words, the card spending amount is a significant factor in determining the customer grade

Scipy Stats module

```
scipy.stats
│
├── 01 T-test                         (one-sample t-test)
│   │
│   ├── ttest_1samp                   (independent-samples t-test)
│   ├── ttest_ind
│   └── ttest_rel                     (paired samples t-test)
│
├── 02 Non parametric test
│   │                                 (Mann Whitney U test - median : almost identical to Wilcoxon rank sum test)
│   ├── mannwhitneyu
│   ├── ranksums                      (Wilcoxon rank sum test - median)
│   └── wilcoxon
│                                     (Wilcoxon signed rank sum test)
├── 03 Normality test
│   │
│   ├── anderson                      (Anderson-Darling , When the number of data is relatively large)
│   ├── kstest                        (Kolmogorov-Smirnov , When the number of data is relatively large)
│   ├── mstats.normaltest
│   └── shapiro                       (shapiro, most stringent normality test, When the number of data is relatively small)
│
├── 04 Equal Variance test
│   │
│   ├── bartlett
│   ├── fligner
│   └── levene
│
├── 05 Chi square test
│   │                                 (Chi-square test of independence, independence test)
│   ├── chi2_contingency              (Chi-square test, goodness-of-fit test)
│   ├── chisquare
│   └── fisher_exact                  (Fisher's exact test, used when the number of cells with a frequency of 5 or less is 20% of the total cells)
│
└── 06 ANOVA                          (one-way analysis of variance)
    │
    └── f_oneway                      (For variance analysis, the statmodels module is better.)
```

Assignment 3

- Submission due : April. 3th, 23:55
- What to submit : Notebook file (.ipynb)
  - Colab : [File]-[Download]-[Download .ipynb]
  - Kaggle : [File]-[Download Notebook]

Not mandatory (No Score)

but if you have available time you could do it by yourself and ask me and TA if you have question.

Assignment 3

In each example, you can see find the function parameter options in Scipy library H.P.
https://docs.scipy.org/doc/scipy/reference/stats.html

You can do
1. Change the shape of the data
2. Draw histogram of the data
3. Change the parameter option of the “scipy.stats” parameter option, and interpret it.
4. **Find an interesting example related with “scipy.stats”, and make it as an ipynotebook with well-formed explanation and submit it. If it’s valuable as I think, I can select as one of mid-term exam sample.**

Q & A

김은희 (ehkim@kisti.re.kr)
TA: Yesim Selcuk (yesimselcuk@kisti.re.kr)